DOCUMENT RESUME

ED 129 904 TM 005 747

AUTHOR Kleinke, David J.
TITLE The Accuracy of Three Approximations for Test Reliability.
PUB DATE Apr 76
NOTE 10p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, California, April 19-23, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Comparative Analysis; Correlation; Standard Error of Measurement; Statistical Analysis; *Test Reliability

ABSTRACT

Data from 200 college-level tests were used to compare three reliability approximations (two of Saupe and one of Cureton) to Kuder-Richardson Formula 20 (KR20). While the approximations correlated highly (about .9) with the reliability estimate, they tended to be underapproximations. The explanation lies in an apparent bias of Lord's approximation to the standard error of measurement. Until further investigation is completed, it is suggested that these approximations be used only for comparisons among tests of similar average relative difficulty. (Author/RC)

* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *

ED129904

THE ACCURACY OF THREE APPROXIMATIONS FOR TEST RELIABILITY

David J. Kleinke
Syracuse University

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Paper presented at meeting of National Council on Measurement in Education; San Francisco; April 22, 1976.

TM005 747

Abstract

Data from two hundred college-level tests were used to compare three reliability approximations (two of Saupe and one of Cureton) to KR20. While the approximations correlated highly (about .9) with the reliability estimate, they tended to be underapproximations. The explanation lies in an apparent bias of Lord's approximation to the standard error of measurement. Until further investigation is completed, it is suggested that these approximations be used only for comparisons among tests of similar average relative difficulty.

The purpose of this study was to investigate three related approximations to a test's reliability. These approximations have been suggested by those who developed them (Cureton and others, 1973 and Saupe, 1961) and by others (e.g., McMorris, 1972 and Payne, 1974) to be sufficiently accurate to warrant their use before, instead of, and/or in addition to computing such estimates as the familiar Kuder-Richardson Formula 20 (KR20).

Each of these approximations is based on Lord's (1957, 1959) approximation to the standard error of measurement,

$$SE_{Meas} \hat{=} \frac{3\sqrt{n}}{7} , \qquad (1)$$

where n is the number of items in a test. If the right-hand side of (1) is set equal to $S_X\sqrt{1-r_{XX'}}$ and solved for $r_{XX'}$,

$$r_{XX'} \hat{=} 1 - \frac{.184n}{S_X^2} . \qquad (2)$$

This result was first noted by Saupe, who also provided an unbaised estimate,

$$r_{XX'} \hat{=} \frac{n}{n-1}\left(1 - \frac{.2n}{S_X^2}\right) . \qquad (3)$$

In both (2) and (3), $r_{XX'}$ is the reliability of the test X, and S_X^2 is its variance.

Much earlier, Jenkins (1946) provided an approximation to the standard deviation,

$$S_X \hat{=} \frac{\Sigma U(1/6) - \Sigma L(1/6)}{.5N} . \qquad (4)$$

$\Sigma U(1/6)$ is the sum of the highest one-sixth of the scores; $\Sigma L(1/6)$ is the sum of the lowest one-sixth; and N is the number of examinees. By squaring

(4) and substituting it for S_X^2 in (2), Cureton and others obtained

$$r_{XX'} \hat{=} 1 - .043n \left(\frac{N}{\Sigma U(1/6) - \Sigma L(1/6)}\right)^2 \quad . \quad (5)$$

Thus there are at least three approximations to the internal consistency of a test, namely Equations (2), (3), and (5). In this study they are designated RS1, RS2, and RC, respectively.

To compare the results of these approximations to KR20, both Pearson correlations and the algebraic and absolute differences between the approximation and KR20 were computed. Also, Equations (2), and (3), and (5) were set equal to KR20 and solved for the coefficeint of n in each case.

Tests and Computations

The tests used were 210 classroom tests for which machine-scorable answer sheets were submitted to Syracuse University's Test Scoring and Evaluation Services for scoring and analysis during a two-month period. Two of the tests were subsequently eliminated from the study when it was discovered that the instructor had submitted blank answer sheets for absent students. The aparent effect of the resulting "scores" of zero on both the variances and the KR20s was marked. Another eight tests were eliminated from most of the study because they had negative KR20s. Selected descriptive statistics for the tests are presented in Table 1.

The answer sheets were scanned on an OpScan DM100 scanner, and routine test analysis statistics, including KR20 and SE_{Meas} based on KR20 were computed using a locally-developed test analysis program. Further computations, specific to this study, were done with subprograms of the Statistical package

Table 1

Test Statistics

Statistic	200 Tests with KR20 greater than zero			208 tests, regardless of KR20		
	Mean	SD	Limits	Mean	SD	Limits
Number of items	44.8	29.8	8 - 140	43.6	29.8	8 - 140
Number of examinees	90.0	104.9	6 - 333	87.1	103.8	5 - 883
Mean	30.4	20.5	4.0 - 102.4	29.5	20.6	2.1 - 102.4
Standard deviation	5.4	3.4	1.3 - 17.4	5.3	3.5	0.7 - 17.4
KR20	.663	.209	.025 - .945	.636	.247	-.144 - .945
Standard error of measurement	2.5	0.9	1.0 - 5.0	2.4	0.9	.7 - 5.0

for the social sciences, (Nie and others, 1975). All data processing was performed on an IBM 370/155 computer.

Results

Correlations among the three approximations and KR20 are presented in Table 2. The values above the diagonal in Table 2 are for all 208 cases.

Table 2

Correlations among KR20 and Three Approximations

	KR20	RS1	RS2	RC
KR20		.862	.847	.876
RS1	.929		.999	.984
RS2	.925-	.999		.981
RC	.916	.982	.980	

Those below the diagonal are for the two hundred cases with KR20 greater than zero. All subsequent data are for these two hundred cases only. The differences

between the approximations and their appropriate statistics are summarized in Table 3, and the coefficients computed by setting Equations (2), (3), and (5) equal to KR20 are presented in Table 4.

Table 3

Summary Statistics for Differences between KR20 and Three Approximations

Difference	Mean	SD	Minimum	Maximum
RS1-KR20				
Algebraic	-.074	.108	-.692	.103
Absolute	.081	.103	.001	.692
RS2-KR20				
Algebraic	-.093	.132	-.822	.059
Absolute	.098	.128	.000	.822
RC-KR20				
Algebraic	-.076	.121	-.687	.131
Absolute	.088	.113	.000	.687

Table 4

Coefficients for Three Approximations

Approximation	Coefficient in Equation	Mean	SD	Minimum	Maximum
RS1	.184	.157	.026	.080	.209
RS2	.200	.169	.028	.088	.228
RC	.043	.037	.007	.019	.051

Discussion

On the evidence in Table 2, the three approximations would be judged to be essentially the same and quite closely related to KR20. That is, these approximations are very highly correlated with each other and also with KR20. However, scrunity of Tables 3 and 4 suggests a systematic difference between each of these approximations and KR20. The approximations tended to yield values that are smaller than the criterion statistic.

Further, the inaccuracy is systematically related to the relative difficulty of the test. All but eleven of the tests had mean item difficulty indices (mean total scores divided by number of items) that were greater than 0.5. The correlations between that variable ($\bar{p}$) and the algebraic differences were -.65, -.60, and -.57 for RS1, RS2, and RC respectively. When the variable $\bar{p}\bar{q} = \bar{p}(1-\bar{p})$ was correlated with these differences, values of .75, .70, and.66 were obtained. That is, the more $\bar{p}$ departed from 0.5, the greater the size of underapproximation.

Correlations between $\bar{p}$ and the coefficients that would have set Equations (2), (3), and (5) equal to KR20 were -.85, -.84, and -.76, respectively. For $\bar{p}\bar{q}$, these correlations rose to .89, .89, and .80. The joint distribution of $\bar{p}$ and the coefficient for RS1 is presented in Table 5. These strong relationships, however, did not carry through to the relationships between relative difficulty and KR20. The correlation between KR20 and $\bar{p}$ was .01; that between $\bar{p}\bar{q}$ and KR20 was.04.

The apparent source of the inaccuracies lies in Equation (1). There were strong and consistent relationships between both $\bar{p}$ and $\bar{p}\bar{q}$ and all measures of the inaccuracies from Equation (1). These are being separately investigated currently.

At this point, it appears that one can safely conclude, however, that there are systematic inaccuracies in Lord's approximation and in approximations based on it. The reliability approximations are nonetheless useful for comparing two tests with approximately the same relative difficulty, for setting a probably lower limit to KR20, and for rough approximations. during the test development process.

As has already been stated, investigation is continuing into the nature, source and generality of these inaccuracies.

Table 5

Scatergram of Coefficient for RS1 and Mean Item Difficulty

CS1---Coefficient for Saupe .184

Mean Item Difficulty	.08	.09	.10	.11	.12	.13	.14	.15	.16	.17	.18	.19	.20	.21
.87-.89	1	1												
.84-.86		1	1	1										
.81-.83			1	6	2	2								
.78-.80				1	8	3	3							
.75-.77					2	6	6	3	1					
.72-.74						5	5	12	5	3				
.69-.71						2	4	5	14	2				
.66-.68						1	0	3	8	8	2	0	1	1
.63-65							1	1	3	6	4	1	0	0
.60-.62								2	2	3	4	2	4	0
.57-.59									1	1	2	4	0	0
.54-.56									0	0	2	3	5	1
.51-.53									2	0	3	3	1	
.48-.50										1	0	1	1	
.45-.47											1	1	1	
.42-.44											1	0		
.39-.41												2		

References

Cureton, E.E., Cook, J.A., Fischer, R.T., Laser, S.A., Rockwell, N.J., & Simmons, J.W. Jr. Length of test and standard error of measurement. Educational and Psychological Measurement, 1973, 33, 63-68.

Jenkins, W.L. A short-cut method for σ and r. Educational and Psychological Measurement, 1946, 6, 533-536.

Lord, R.M. Do tests of the same length have the same standard errors of measurement? Educational and Psychological Measurement, 1957, 17, 510-521.

Lord, F.M. Tests of the same length do have the same standard error of measurement. Educational and Psychological Measurement, 1959, 19, 233-239.

McMorris, R.F. Evidence on the quality of several approximations for commonly used measurement statistics. Journal of Educational Measurement, 1972, 9, 113-122.

Nie, N.H., Hull, C.H., Jenkins, J.G., Steinbrenner, K., & Bent, D.H. Statistical package for the social sciences (2nd ed.). New York: McGraw-Hill, 1975.

Payne, D.A. The assessment of learning: Cognitive and affective. Lexington, MA: D.C. Heath, 1974.

Saupe, J.L. Some useful estimates of the Kuder-Richardson Formula number 20 reliability coefficente. Educational and Psychological Measurement, 1961, 21, 63-71.